TRAINING SET OPTIMISATION FOR LINEAR B-CELL EPITOPE PREDICTION

Author: MohammadJavad AghababaieBeni

Program: MSc Computer Science

Supervisor: Dr Roberto Alamino

OCTOBER 2024

ASTON UNIVERSITY OF BIRMINGHAM

Table of Contents

Declaration:

I declare that I have personally prepared this assignment. The work is my own, carried out personally by me unless otherwise stated and has not been generated using paid for assessment writing services or Artificial Intelligence tools unless specified as a clearly stated approved component of the assessment brief. All sources of information, including quotations, are acknowledged by means of the appropriate citations and references. I declare that this work has not gained credit previously for another module at this or another University, save for permitted elements which formed part of an associated proposal linked directly to this submission.

I understand that plagiarism, collusion, copying another student and commissioning (which for the avoidance of doubt includes the use of essay mills and other paid for assessment writing services, as well as unattributed use of work generated by Artificial Intelligence tools) are regarded as offences against the University's Assessment Regulations and may result in formal disciplinary proceedings.
I understand that by submitting this assessment, I declare myself fit to be able to undertake the assessment and accept the outcome of the assessment as valid.

Student signature: MohammadJavad AghababaieBeni

Date: 13 October 2024

Acknowledgements

I would like to express my deepest gratitude to my academic supervisor, Dr. Roberto Alamino, for his invaluable guidance, support, and encouragement throughout the course of this research. His expertise and insights were crucial to the success of this project.

I am also grateful to the Department of Computer Science for providing the necessary resources, including cloud servers, that allowed me to carry out the computational aspects of this research.

A special thanks goes to Dr. Felipe Campelo, my previous supervisor, who first introduced me to the world of Artificial Intelligence and Data Science, setting the foundation for my journey in this field.

would like to thank Mr Onawole and Dr. Ashford who worked on the same research area at Aston University for providing invaluable insights and data related to epitope prediction models, which greatly contributed to this research.

I would also like to extend my heartfelt thanks to my family for their continuous support and encouragement throughout this journey. In particular, I am deeply thankful to my mother, whose unwavering belief in me has been a constant source of strength and motivation. Her love and encouragement have been vital in helping me complete this research.

Abstract

This dissertation investigates the prediction of linear B-cell epitopes (LBCEs) with a focus on organism-specific models, using advanced machine learning techniques. The research primarily aims to assess the effectiveness of models trained on pathogen-specific datasets compared to generalist models, with an emphasis on applications for the Coronavirus. A comprehensive dataset was curated from the Immune Epitope Database (IEDB), comprising various pathogens, with a focus on Coronavirus for specific epitope prediction. The study employed feature selection techniques including Boruta and Genetic Algorithms (GA) to refine and optimize the feature set, significantly reducing dimensionality while preserving predictive power.

Two machine learning models—Feedforward Neural Network (FNN) and XGBoost—were developed and evaluated based on their ability to predict LBCEs. XGBoost outperformed FNN in both organism-specific and heterogeneous datasets, demonstrating superior metrics such as Area Under the Curve (AUC), F1 score, and Matthews Correlation Coefficient (MCC). The research also explored the generalization capabilities of XGBoost across multiple pathogens, confirming its broader applicability in immunoinformatics and vaccine development.

Key findings underscore the importance of organism-specific training for improving prediction accuracy, while hybrid and ensemble approaches are recommended for further enhancing generalizability. The dissertation also addresses challenges related to class imbalance and computational efficiency by employing techniques such as SMOTE, Focal Loss, and cloud-based resources for model training.

This study contributes to the field of epitope prediction by offering practical insights for developing predictive models that can generalize across pathogens, facilitating more efficient vaccine design and therapeutic antibody development.

Introduction

The human immune system is a sophisticated defence mechanism that identifies and neutralizes foreign invaders such as bacteria, viruses, and parasites. A crucial aspect of this defence is the recognition of epitopes, specific regions on antigens that are identified by immune cells to initiate a targeted immune response (Sette & Fikes, 2003). Among these epitopes, B-cell epitopes play an essential role in humoral immunity, which involves the production of antibodies by B-cells to combat infections (El-Manzalawy & Honavar, 2010). The accurate identification of B-cell epitopes is fundamental for several medical applications, including vaccine development, therapeutic antibody design, and disease diagnostics (SoriaGuerra et al., 2015).

B-cell epitopes can be classified into linear and conformational epitopes. Linear B-cell epitopes (LBCEs) consist of continuous sequences of amino acids, whereas conformational epitopes involve amino acids that may be far apart in the primary sequence but are brought together in the protein's three-dimensional folded structure (Punt et al., 2018). Due to their relative stability and ease of synthesis, LBCEs are often prioritized in computational epitope prediction, making them ideal candidates for vaccine design, particularly for peptide-based vaccines (Malik et al., 2022).

Traditional methods for identifying B-cell epitopes, such as enzyme-linked immunosorbent assay (ELISA), phage display, and X-ray crystallography, provide highly accurate data but are resource-intensive, time-consuming, and impractical for large-scale screening (Yang & Yu, 2009). To overcome these challenges, computational or in silico methods have been developed as a more scalable and cost-effective alternative (Chen et al., 2007). Computational prediction methods can be categorized broadly into sequence-based and structure-based approaches. Sequence-based approaches utilize features derived from the primary structure, such as hydrophilicity, amino acid composition, and antigenicity to predict LBCEs. These methods are computationally efficient but may fail to capture complex biological interactions that require spatial information (Jespersen et al., 2017). On the other hand, structure-based approaches rely on the three-dimensional structure of proteins, providing a more accurate representation of conformational epitopes but requiring high-quality 3D data, which limits their widespread application (Isidro et al., 2015; Yao et al., 2013).

A recent trend in epitope prediction has been the use of machine learning (ML) models, which have demonstrated considerable success in improving predictive accuracy by learning complex data patterns (Cia et al., 2023). Models such as support vector machines (SVMs), random forests (RFs), and deep neural networks have been widely adopted for LBCE prediction, offering significant improvements over traditional computational approaches (Soria-Guerra et al., 2015). The evolution of tools like BepiPred—from using hidden Markov models to adopting deep learning techniques such as protein language models—has further enhanced the field (Jespersen et al., 2017; Clifford et al., 2022).

The main objective of this dissertation is to explore the benefits of organism-specific training for predicting LBCEs, with a specific focus on Coronavirus. This study aims to determine whether models trained on pathogen-specific datasets can outperform generalist models when predicting LBCEs for the target organism. In addition, this research seeks to develop a generalized approach that performs well across various datasets. By leveraging advanced ML techniques, such as feature selection using

Genetic Algorithms (GAs), the study aims to balance organism-specific accuracy with cross-pathogen applicability, ultimately contributing to a more flexible and reliable approach to epitope prediction.

Literature Review

Introduction to Epitope Prediction

The prediction of epitopes forms an essential component of immunology, serving as the basis for comprehending how immune responses are triggered to identify and eliminate pathogens. Specifically, B-cell epitopes are sections of antigens that are detected by the immune system, initiating the production of antibodies and prompting humoral immunity (Sette & Fikes, 2003). This recognition process is fundamental for the activation of humoral immunity, which plays a critical role in combating infections through antibody production (El-Manzalawy & Honavar, 2010). The accurate identification of epitopes has widespread applications in vaccine development, therapeutic antibody design, and diagnostic tools (Soria-Guerra et al., 2015).

B-cell epitopes are categorized into two main types: linear and conformational. Linear epitopes comprise continuous stretches of amino acids, while conformational epitopes involve residues that are separated in the primary sequence but come together upon protein folding to form functional epitopic regions (Punt et al., 2018). The focus of most computational research has been on linear B-cell epitopes (LBCEs), owing to their predictable nature and the availability of extensive sequence data (Malik et al., 2022). In contrast, conformational epitopes, while important, are far more complex due to their reliance on the protein's tertiary structure, which is often difficult to predict accurately (Yang & Yu, 2009).

Traditional Methods for Epitope Identification

Techniques such as enzyme-linked immunosorbent assay (ELISA), X-ray crystallography, nuclear magnetic resonance (NMR), and phage display have traditionally served as core methods for identifying epitopes (Yang & Yu, 2009).. These techniques, while accurate, are resource-intensive, expensive, and require a significant amount of time and expertise, making them impractical for high-throughput screening (Soria-Guerra et al., 2015). For instance, Xray crystallography is highly effective for resolving the structures of antibody-antigen complexes, but the requirement for crystallization is often a major bottleneck. Similarly, NMR spectroscopy provides valuable information regarding epitope structure but is limited to smaller proteins and requires substantial resources (Ponomarenko & Bourne, 2008).

Given these limitations, there has been an increasing emphasis on the development of computational (in silico) approaches for epitope prediction. These methods offer a more scalable and cost-effective solution, enabling rapid identification of potential epitopes and allowing researchers to prioritize experimental validation more efficiently (Chen et al., 2007).

Computational Approaches to Epitope Prediction

Computational epitope prediction methods can be divided into sequence-based and structure-based approaches.

Sequence-Based Approaches

Sequence-based methods rely on analyzing the primary structure of proteins, using features such as amino acid composition, hydrophilicity, polarity, and antigenicity. These methods have been effective in predicting LBCEs, as they leverage accessible sequence data and provide computational efficiency for rapid analysis (Chen et al., 2007; Jespersen et al., 2017).

The hydrophilicity of amino acids is a particularly important feature for LBCE prediction, as hydrophilic residues are more likely to be located on the surface of the protein, making them accessible to antibodies (Pellequer et al., 1993). Early prediction models used the Parker hydrophilicity scale and Chou-Fasman beta-turn analysis to predict epitope locations, emphasizing regions with high surface exposure and flexibility (Pellequer et al., 1993; Saha & Raghava, 2006).

BepiPred, one of the most widely used tools for sequence-based epitope prediction, has undergone multiple iterations:

- BepiPred-1.0 utilized hidden Markov models (HMM), which were effective for identifying linear epitopes based on sequence motifs but limited by their inability to capture complex relationships between distant residues (Larsen et al., 2006).
- BepiPred-2.0 integrated random forest (RF) algorithms, combining features such as antigenicity and surface exposure, resulting in enhanced prediction performance and robustness (Jespersen et al., 2017).
- The latest version, BepiPred-3.0, adopts deep learning techniques, specifically protein language models, to learn complex dependencies and improve the accuracy of LBCE prediction (Clifford et al., 2022).

Structure-Based Approaches

Structure-based approaches take advantage of the 3D conformation of proteins to predict epitopes. These methods provide a more accurate prediction for conformational epitopes by evaluating factors such as solvent accessibility, secondary structure, and tertiary interactions (Isidro et al., 2015). Tools like ElliPro and Discotope utilize protein structures from databases such as the Protein Data Bank (PDB) to identify regions likely to form epitopes based on their surface location and structural properties (Sussman et al., 1998; Ponomarenko et al., 2008).

However, structure-based methods are limited by the availability of high-quality protein structural data. While methods like homology modeling can generate structural models, the reliability of these models varies based on the similarity between the target protein and known structures, thus affecting prediction accuracy (Yao et al., 2013). 4. Machine Learning Techniques in Epitope Prediction

The advent of machine learning (ML) has significantly advanced the field of epitope prediction by enabling the analysis of complex relationships between sequence and structural features. Support Vector Machines (SVMs), Random Forests (RFs), and deep learning models have all contributed to improving the predictive accuracy of LBCE identification (Cia et al., 2023).

- Support Vector Machines (SVMs): SVMs are highly effective for classification tasks in LBCE prediction due to their ability to work well with high-dimensional feature spaces, such as those encountered in proteomics. LBtope, for example, uses an SVM-based approach, trained on experimentally validated epitopes, to improve prediction specificity and minimize over-prediction (Singh et al., 2013).
- Neural Networks: The use of neural networks, particularly feed-forward neural networks (FFNNs), has introduced a layer of complexity capable of capturing nonlinear relationships in data. ABCpred, which uses an artificial neural network for LBCE prediction, demonstrated improvements over traditional linear models, showing the potential for neural networks to handle the complex nature of epitope prediction (Saha & Raghava, 2006).
- Deep Learning Models: More recent advancements include the application of deep learning (DL). Models such as BepiPred-3.0 use protein language embeddings to capture sequence features that contribute to epitope recognition. These models leverage data from large-scale protein sequences, enabling the extraction of nuanced features beyond what traditional ML models can achieve (Clifford et al., 2022). EpiDope, which uses Long Short-Term Memory (LSTM) networks, can model dependencies within protein sequences, providing insights into epitope prediction that conventional sequence-based methods may miss.

Organism-Specific Training for LBCE Prediction

Organism-specific training has emerged as a promising approach to address the limitations of generalist epitope prediction models. By tailoring the training dataset to epitopes from a specific pathogen, organism-specific models aim to enhance the prediction accuracy for that particular organism (Ashford, 2023).

Advantages of Organism-Specific Models

The specificity of training data allows organism-specific models to focus on unique features of the pathogen. Studies have shown that organism-specific models outperform generalist models when applied to the target organism, achieving higher metrics such as precision, recall, and Matthews Correlation Coefficient (MCC) (Onawole, 2023). This approach has been particularly effective for

Coronavirus, where distinct sequence motifs and unique structural features necessitate specialized training to capture immunogenic regions accurately (Ashford et al., 2023).

Challenges and Strategies to Address Generalizability

The main limitation of organism-specific models lies in their generalizability. These models, while highly accurate for the organism they are trained on, often fail to perform well on unrelated pathogens due to their specificity. This lack of versatility presents a challenge for broad-spectrum vaccine and therapeutic development (Singh et al., 2013).

To address this, recent research has explored the use of hybrid models, which incorporate both organism-specific and cross-pathogen features. Transfer learning and meta-learning are also being investigated as strategies to improve generalizability by enabling models to adapt to new pathogens with minimal retraining, thus making them more suitable for emerging infectious diseases (Ashford et al., 2023).

Feature Engineering and Optimization Techniques

Feature engineering is critical to improving the predictive power of machine learning models in epitope prediction. The selection of relevant features not only influences model accuracy but also affects computational efficiency and generalizability.

Feature Extraction and Selection

Feature extraction involves identifying biologically significant properties from protein sequences and structures. These features include amino acid composition, hydrophilicity, surface accessibility, polarity, and sequence motifs (Katoch et al., 2021). Given the high dimensional nature of biological data, feature selection techniques are essential for refining the dataset to include only the most predictive elements, thereby reducing overfitting and enhancing model interpretability.

Principal Component Analysis (PCA) has been widely adopted for dimensionality reduction, helping to retain the most informative components while discarding redundant data. Autoencoders, a type of unsupervised deep learning model, are also used for dimensionality reduction by learning efficient compressed representations of the input features (Shukla et al., 2015). These techniques play a crucial role in handling high-dimensional data typical in epitope prediction.

Genetic Algorithms (GAs) and Hybrid Feature Selection

Genetic Algorithms (GAs) are powerful tools for optimizing feature sets in LBCE prediction. Inspired by natural evolutionary processes, GAs iteratively refine feature subsets based on their "fitness," which is measured by their contribution to model accuracy (Katoch et al., 2021). This method allows for the identification of feature combinations that maximize predictive power while minimizing unnecessary complexity.

The Boruta algorithm, often used in tandem with Random Forests (RFs), is another effective technique for determining feature importance. Boruta eliminates irrelevant features in a statistically sound manner, ensuring that only the most significant variables are retained, thus improving both precision and recall in LBCE prediction models (Ashford, 2023). The combination of PCA, GAs, and Boruta forms a hybrid feature selection approach that leverages the strengths of each method, resulting in a more streamlined and efficient feature set.

Meta-Features and Deep Learning

Recent research has focused on developing meta-features—higher-order features derived from primary data that encapsulate complex biological information. Deep learning models, such as convolutional neural networks (CNNs) and LSTMs, have been instrumental in learning these meta-features directly from raw sequence and structural data. For instance, CNNs are adept at capturing spatial hierarchies in data, which is particularly useful for understanding the local and global dependencies within protein sequences (Clifford et al., 2022).

The use of transfer learning in deep learning has allowed models to leverage pretrained networks trained on large protein datasets. This approach enables models to incorporate previously learned features, significantly improving their ability to predict epitopes in novel contexts (Vaswani et al., 2017). The use of pretrained embeddings, such as those from protein language models, has made it possible to capture subtle relationships in protein sequences that would otherwise be missed using traditional feature extraction techniques.

Recent Trends in Epitope Prediction Research

The landscape of epitope prediction has evolved significantly with advances in computational biology. Deep learning techniques, such as transformer models and protein embeddings, have set new benchmarks in terms of predictive accuracy and generalizability (Vaswani et al., 2017). These models leverage large protein datasets, such as UniProt and PDB, to learn generalized features that are applicable across a wide variety of pathogens (Sussman et al., 1998).

Transformer-Based Models

Transformers, originally developed for natural language processing, have been adapted to handle the complexities of protein sequence analysis. These models employ a self-attention mechanism to capture long-range dependencies within sequences, enabling more precise identification of potential epitopes (Vaswani et al., 2017). ProtBERT, a transformer model trained specifically on protein data, has been used to predict B-cell epitopes by representing protein sequences in a manner that highlights key antigenic regions (Clifford et al., 2022).

Ensemble Learning and Hybrid Models

Ensemble learning has become an increasingly popular strategy to enhance the robustness of epitope prediction models. By combining the predictions of multiple models—such as random forests, deep learning networks, and SVMs—ensemble approaches reduce the likelihood of overfitting and improve generalization across diverse datasets (Cia et al., 2023). For instance, ensemble models that incorporate both sequence-based and structure-based methods have demonstrated improved performance by leveraging the strengths of each approach, providing a more comprehensive predictive framework.

The integration of hybrid models, which combine traditional machine learning, deep learning, and meta-learning techniques, has also proven effective in adapting to new and evolving pathogens. This approach allows for rapid model adaptation, which is particularly valuable in pandemic scenarios where new variants of a virus may emerge (Ashford et al., 2023).

General Applications in Vaccine and Therapeutic An)body Development

The application of LBCE prediction in vaccine development and therapeutic antibody design is one of the most impactful outcomes of this research area.

Vaccine Development

Linear B-cell epitopes are highly suitable for use in peptide vaccines due to their stability and ability to elicit strong immune responses (Punt et al., 2018). The identification of LBCEs on viral antigens allows for the design of vaccines that specifically target these regions, thereby inducing protective immunity. This approach has been utilized in the development of vaccines for influenza, HIV, and more recently, SARS-CoV-2 (Onawole, 2023). Computational prediction tools played a pivotal role in accelerating vaccine development during the COVID19 pandemic, enabling the identification of immunogenic regions on the spike protein of the virus (Ashford et al., 2023).

Therapeutic Antibody Design

In therapeutic antibody design, the identification of neutralizing epitopes is crucial for developing antibodies that can effectively prevent infection. Monoclonal antibodies that bind to these neutralizing epitopes can block critical viral functions, such as host cell entry. The use of in silico epitope prediction has expedited the process of identifying these target sites, enabling more efficient development of therapies for viral infections like COVID-19 and Ebola (El-Manzalawy & Honavar, 2010).

Limitations and Challenges in Current Approaches

Despite significant advancements, several challenges continue to limit the effectiveness of current epitope prediction approaches.

Data Quality and Availability

The quality of training data is a critical determinant of the performance of computational models. Many available datasets are derived from experiments that vary in accuracy, and the lack of standardization across different datasets can introduce biases that affect model reliability (Vita et al., 2008). Data imbalance, where certain pathogens or epitope types are overrepresented, further complicates the training process, leading to models that perform poorly on underrepresented targets (Singh et al., 2013).

Predicting Conformational Epitopes

The prediction of conformational epitopes remains a significant challenge due to the reliance on high-resolution 3D protein structures, which are often unavailable. Although homology modeling and Cryo-EM have improved the accessibility of structural data, these methods are still not practical for high-throughput applications, limiting the broad application of structure-based approaches (Yao et al., 2013).

Generalizability and Overfitting

One of the biggest challenges is balancing model specificity and generalizability. While organism-specific models can achieve high accuracy for their target organism, they often suffer from overfitting and fail to generalize well to other pathogens. This issue becomes particularly problematic in situations where rapid adaptation to new pathogens is required, such as during emerging outbreaks (Ashford, 2023).

Addressing Challenges Through Novel Approaches

Cross-Pathogen Feature Generalization

To enhance generalizability, cross-pathogen feature generalization has been proposed as a solution. This involves training models using features that are common across multiple pathogens, which helps in developing models that retain predictive power even when applied to new organisms (Onawole, 2023).

Meta-Learning and Transfer Learning

Transfer learning has shown great promise in improving epitope prediction for pathogens with limited data. For instance, ProtBERT and ProtTrans, transformer-based models trained on billions of protein sequences, are used to generate embeddings that can be fine-tuned for specific tasks like epitope prediction. These pretrained embeddings have significantly improved performance in LBCE identification by allowing the model to utilize features that capture a wide range of sequence and structural properties, thus enhancing its generalizability (Vaswani et al., 2017; Clifford et al., 2022).

Ensemble and Hybrid Approaches for Robust Predictions

Ensemble learning is another effective method to improve robustness and reduce overfitting in LBCE prediction models. By combining different types of models, such as support vector machines (SVMs), random forests (RFs), and deep learning models like convolutional neural networks (CNNs) and LSTMs, ensemble approaches can provide a balanced and comprehensive predictive outcome. This method leverages the complementary strengths of different algorithms to achieve more accurate predictions, making the final model less sensitive to the weaknesses of any individual model (Cia et al., 2023).

Hybrid models that combine both sequence-based and structure-based features have also been proposed. These models incorporate features such as amino acid composition, solvent accessibility, and secondary structural elements to create a holistic representation of the protein (Jespersen et al., 2017). By integrating data from multiple domains, hybrid approaches ensure that both local sequence characteristics and broader structural information are used for epitope prediction, resulting in a more robust predictive model.

Generative Models and Synthetic Data Augmentation

One of the newer trends in computational biology is the use of generative models to create synthetic data for training purposes. Generative Adversarial Networks (GANs) can be used to generate synthetic epitope sequences that mimic real epitopes, thereby augmenting training datasets and addressing the issue of data scarcity (Goodfellow et al., 2014). Synthetic data augmentation can enhance the model's ability to generalize by exposing it to a broader array of possible epitopes, effectively reducing overfitting and improving robustness.

Variational Autoencoders (VAEs) are also employed to generate novel protein sequences that share structural and functional similarities with known epitopes. This approach can potentially be used to explore novel vaccine candidates, particularly for emerging pathogens for which limited data is available (Kingma & Welling, 2013).

Case Studies in COVID-19 Vaccine and Therapeutic Development

Vaccine Development: A Case Study of COVID-19

The global COVID-19 pandemic has underscored the importance of rapid vaccine development and highlighted the role of computational prediction in accelerating this process. Predictive models were used to identify the spike protein epitopes of SARS-CoV-2, which played a critical role in the development of mRNA vaccines by companies like Pfizer and Moderna (Ashford et al., 2023). The ability to quickly identify linear epitopes that could elicit a strong immune response was essential in developing these vaccines within a short timeframe.

Additionally, the application of organism-specific training in the context of COVID-19 showed that models trained on coronavirus-specific epitopes were more effective at predicting immunogenic regions than generalist models. This specificity allowed for a more focused vaccine design, ultimately improving the vaccine's efficacy by targeting the most immunogenic portions of the virus (Onawole, 2023).

Therapeutic Antibody Design and Monoclonal Antibodies

The identification of neutralizing epitopes has been pivotal in the development of monoclonal antibody therapies for treating viral infections. Neutralizing epitopes are regions on the pathogen where antibody binding can inhibit functions such as viral entry or replication. The use of in silico models to predict these epitopes enables the rapid development of therapeutic antibodies for diseases such as COVID-19 and HIV (El-Manzalawy & Honavar, 2010).

Monoclonal antibodies like Regeneron's REGN-COV2, which were developed to target SARS-CoV-2, benefited from computational epitope prediction to identify effective binding sites on the virus (Malik et al., 2022). This process greatly reduced the timeline for development compared to traditional antibody discovery methods, which often rely on labor intensive laboratory screening.

Summary of Literature Review

The Literature Review provides a comprehensive examination of the field of epitope prediction, highlighting the progression from traditional experimental methods to sophisticated computational approaches that leverage machine learning and deep learning. Early efforts focused on sequence-based and structure-based methods that utilized fundamental biological features like hydrophilicity, amino acid composition, and surface accessibility. These methods provided the groundwork for more advanced computational models, although their efficacy was often limited by data availability and complexity in protein folding.

Machine learning models, including support vector machines (SVMs), random forests (RFs), and neural networks, have significantly improved the predictive accuracy of LBCEs by analyzing complex sequence and structural relationships. Tools like BepiPred, LBtope, and ABCpred have utilized ML techniques to enhance specificity and accuracy, setting a precedent for further development in the field (Saha & Raghava, 2006; Singh et al., 2013; Jespersen et al., 2017).

The adoption of deep learning has brought about major advancements, allowing for the extraction of nuanced features through models like LSTM and transformers (Clifford et al., 2022; Vaswani et al., 2017). These techniques, combined with transfer learning and meta learning, have addressed some of the challenges related to data scarcity and generalizability, especially for pathogens with limited available data.

Organism-specific training has demonstrated significant benefits for predicting LBCEs in targeted pathogens, such as SARS-CoV-2, by focusing on pathogen-specific features that enhance prediction accuracy (Ashford et al., 2023). However, the main limitation of this approach is the model's ability to generalize to new or unrelated pathogens. To address these issues, the development of hybrid and ensemble learning models has been proposed to improve robustness while retaining specificity.

Advanced feature engineering techniques such as Genetic Algorithms (GAs), Boruta, and hybrid feature selection methods have further improved model performance by optimizing feature sets and reducing dimensionality, ensuring that only the most informative features are used during model training (Katoch et al., 2021; Shukla et al., 2015). The use of generative models such as GANs and VAEs represents a frontier in the creation of synthetic data, allowing researchers to augment training datasets and reduce biases due to data limitations (Goodfellow et al., 2014; Kingma & Welling, 2013).

The application of epitope prediction extends far beyond academic research, having real world implications in vaccine design and therapeutic antibody development. During the COVID19 pandemic, the use of computational tools enabled rapid identification of immunogenic epitopes, which were instrumental in the development of effective vaccines and antibody therapies. These advances underscore the critical importance of computational prediction models in responding to emerging global health threats efficiently and effectively (Onawole, 2023; Malik et al., 2022).

In conclusion, while substantial progress has been made in epitope prediction, challenges such as data quality, generalizability, and the complexity of predicting conformational epitopes remain. Addressing these challenges requires a combination of innovative feature selection, deep learning techniques, and a focus on cross-pathogen adaptability. This dissertation aims to contribute to these ongoing efforts by exploring novel organism-specific training methods while ensuring versatility and broad applicability through the use of advanced computational approaches.

Methodology

Dataset

The preparation of an appropriate dataset is essential for ensuring the accuracy and integrity of any research study. In the current study, an extensive heterogeneous dataset was curated, targeting multiple pathogens such as Flu, Epstein-Barr virus, Hepatitis, Lentivirus, and others. This dataset was compiled from the comprehensive XML export of the Immune Epitope Database (IEDB), with a particular focus on obtaining representative epitopes and non-epitopes relevant to immunoinformatics studies.

The complete dataset comprises 601,192 samples, each consisting of 393 features capturing various biochemical and structural properties of amino acids. These features were derived from the ESM-1b protein feature model [Rives et al., 2021]. ESM-1b is a large-scale transformer-based language model trained on over 250 million protein sequences. It captures both evolutionary and structural information about proteins by encoding their sequences into high-dimensional representations. This makes ESM-1b particularly useful for tasks such as epitope prediction, where both sequence and structural properties play a crucial role. By leveraging this model, we ensure that the dataset incorporates detailed biochemical and structural features of the peptides, which are essential for accurate epitope prediction. The dataset underwent preprocessing to ensure data quality—entries with inconsistent or missing information, specifically in relation to protein IDs or peptide positional data, were meticulously removed.

For the purpose of this study, a subset focusing specifically on the coronavirus was extracted from the main heterogeneous dataset. This extraction was performed using the taxonomic ID of the coronavirus, particularly focusing on protein ID 290028, as verified through the NCBI database. The coronavirus subset contains a refined set of samples, where only linear B-cell epitopes were selected, with peptide lengths restricted to between 8 and 25 amino acids. This range ensures that overly short or extended sequences are excluded, thus reducing noise and redundancy in the data.

The coronavirus-specific dataset was subsequently divided into training, validation, and holdout sets, enabling robust model development and evaluation. This focused approach allows for a detailed investigation into coronavirus-specific immune response prediction, while the broader heterogeneous dataset serves as a basis for potential future studies involving multiple pathogens.

The methodology for this research is structured to address the complexity of predicting coronavirus-specific epitopes using a variety of machine learning techniques and computational tools. This approach consists of several critical components: feature engineering, model development, handling class imbalance, hyperparameter tuning, and model evaluation, with a focus on utilizing cloud-based computational resources to overcome practical challenges. Each aspect is explained in detail below to highlight its significance, implementation, and contribution to achieving accurate and reliable predictions.

Feature Engineering: Boruta and Genetic Algorithm (GA)

The coronavirus dataset used in this study comprises 393 features, each capturing distinct biochemical properties of amino acids. Given the high dimensionality of this dataset, effective feature selection becomes imperative to improve model interpretability, reduce computational costs, and avoid overfitting. The feature engineering process consists of two major steps: **Boruta Feature Selection** and **Genetic Algorithm (GA)** optimization.

Boruta Feature Selection

Boruta is a robust all-relevant feature selection method based on the Random Forest algorithm, designed to identify and retain features that have a statistically significant impact on the target variable

(Kursa & Rudnicki, 2010). In this research, Boruta was used to rank and evaluate the importance of each feature in predicting whether an amino acid segment functions as an epitope.

The method operates by creating shadow features, which are randomized duplicates of the original features. These shadow features serve as a baseline to compare the importance of the actual features. Boruta iteratively evaluates the relevance of each feature by comparing it with its shadow counterpart. Features that consistently show higher importance than their shuffled shadows are considered relevant and retained. Conversely, those that perform worse than the shadows are removed. This process ensures that only the most informative and meaningful features are kept, while irrelevant or redundant data is filtered out.

The application of Boruta in this study was particularly effective in reducing the feature set from 393 to a more manageable subset. This refinement was crucial for improving model interpretability and reducing computational complexity. By eliminating noisy and non-informative features, the risk of overfitting, particularly with a complex dataset, was minimized. This careful feature selection laid a solid foundation for subsequent optimization using Genetic Algorithms, ensuring that the most relevant features were used in further steps of the predictive model development.

Genetic Algorithm (GA) for Feature Subset Refinement

After applying Boruta, the next step involved **Genetic Algorithm (GA)** to further refine the selected feature subset. **GA** is a heuristic search algorithm inspired by the principles of natural selection and genetics. It iteratively evolves candidate solutions toward optimal or near optimal subsets of features by mimicking biological processes such as selection, crossover, and mutation.

The motivation behind using GA in this context was twofold: (1) to optimize the feature subset for better model performance and generalizability, and (2) to balance the dataset for enhanced classification of the minority class (epitopes). Unlike traditional deterministic approaches that may get stuck in local optima, GA's stochastic nature allows it to explore a wider search space, thus effectively preventing premature convergence and ensuring a more comprehensive evaluation of possible feature combinations.

Detailed GA Workflow

1. **Initialization**: The process begins by generating an initial population of candidate solutions. Each candidate, often referred to as a "chromosome," represents a unique subset of features encoded as a binary vector, where each bit indicates whether a corresponding feature is included (1) or excluded (0).
2. **Fitness Function Evaluation**: Each chromosome's quality is evaluated using a **fitness function**, which, in this study, measures the performance of a simple classifier (e.g., a Decision Tree) trained using the corresponding feature subset. The evaluation metric employed in the fitness function typically involves a balance between accuracy and model simplicity—models that yield high performance on the validation set while maintaining lower complexity are scored higher.

3. **Selection**: The selection process involves choosing chromosomes based on their fitness scores. This selection is akin to "survival of the fittest," where individuals with higher fitness have a higher probability of passing their genetic material to the next generation. Methods such as **roulette wheel selection** or **tournament selection** were employed to ensure diversity while focusing on promising solutions.

4. **Crossover and Mutation**: To introduce variability and explore new feature combinations, GA uses crossover (recombination of two parent chromosomes) and mutation (random flipping of bits in a chromosome):

    - **Crossover**: Two selected parent chromosomes exchange parts of their genes, resulting in two offspring that inherit features from both parents. The crossover rate determines how often crossover happens, and it is typically set to ensure sufficient diversity.

    - **Mutation**: To prevent the algorithm from getting stuck in local optima, mutation is applied at a low probability. By randomly flipping some bits, mutation introduces new features into the population, helping the algorithm explore previously unconsidered feature subsets.

5. **Evolution and Convergence**: This process of evaluation, selection, crossover, and mutation repeats for several generations. The algorithm gradually converges toward an optimal subset of features that maximizes the fitness function. The final subset not only retains the features that best predict epitope presence but also ensures that the model is robust and generalizable.

GA's Role in Handling Class Imbalance

Class imbalance—where the number of non-epitope instances vastly outnumbers the epitopes—poses a significant challenge in predictive modeling. GA contributes to mitigating this issue in the following ways:

- **Balanced Sampling**: GA was configured to focus on creating subsets of data that promote balanced representation across classes. By prioritizing samples from the minority class during the feature selection process, GA ensures that the final model receives a more balanced training set.

- **Diverse Representation**: The evolutionary strategy of GA emphasizes diversity. By selecting features that lead to higher sensitivity toward the minority class, GA prevents the model from learning a biased decision boundary, thus reducing false negatives and enhancing the recall for epitopes.

The synergistic use of Boruta and GA provides an efficient means of navigating the high dimensional feature space while ensuring that the models are trained on a well-representative subset of the data. This two-step approach—Boruta for initial filtering and GA for optimization—results in a feature set that is both manageable in size and highly informative, paving the way for effective model training.

Model Development: Feedforward Neural Network (FNN) and XGBoost

Following feature selection, two machine learning models—**Feedforward Neural Network (FNN)** and **XGBoost**—were developed to explore different modeling paradigms suitable for the task at hand.

Feedforward Neural Network (FNN)

The **Feedforward Neural Network (FNN)** is a type of artificial neural network where information moves in only one direction—from the input layer, through hidden layers, to the output layer. This model was chosen for its ability to model complex, non-linear interactions among the selected features, which include biochemical and structural properties of amino acids.

FNN Architecture

Input Layer: The input layer consists of 88 nodes, corresponding to the features selected after Genetic Algorithm (GA) optimization. This dimensionality reduction significantly improves computational efficiency compared to the original dataset, which contained 393 features.

Hidden Layers: The model consists of two hidden layers:

The first hidden layer contains 92 units, each fully connected (Dense layer), with a ReLU activation function.

The second hidden layer contains 38 units, also fully connected, and uses the ReLU activation function to introduce non-linearity, allowing the network to model complex relationships in the data.

Dropout Layers: To prevent overfitting, Dropout layers are included after each hidden layer with a dropout rate of 0.34. Dropout deactivates 34% of the nodes during each training iteration, which encourages the model to learn more robust feature representations by preventing reliance on specific nodes.

Output Layer: The output layer contains a single node with a sigmoid activation function, producing a probability score that predicts whether the input sequence is an epitope. The output is a value between 0 and 1, representing the likelihood of the input being classified as an epitope.

Hyperparameters

Batch Size: The model was trained with a batch size of 14, meaning that the network processes 14 training samples before updating its weights.

Learning Rate: The learning rate was set to 0.0002, determined via hyperparameter optimization to balance the trade-off between convergence speed and model stability.

Epochs: The model was trained for 20 epochs, which represents 20 complete passes over the training dataset.

Best Threshold: The optimal threshold for converting the predicted probability into a binary classification was found to be 0.68. Predictions with a score higher than this threshold were classified as epitopes.

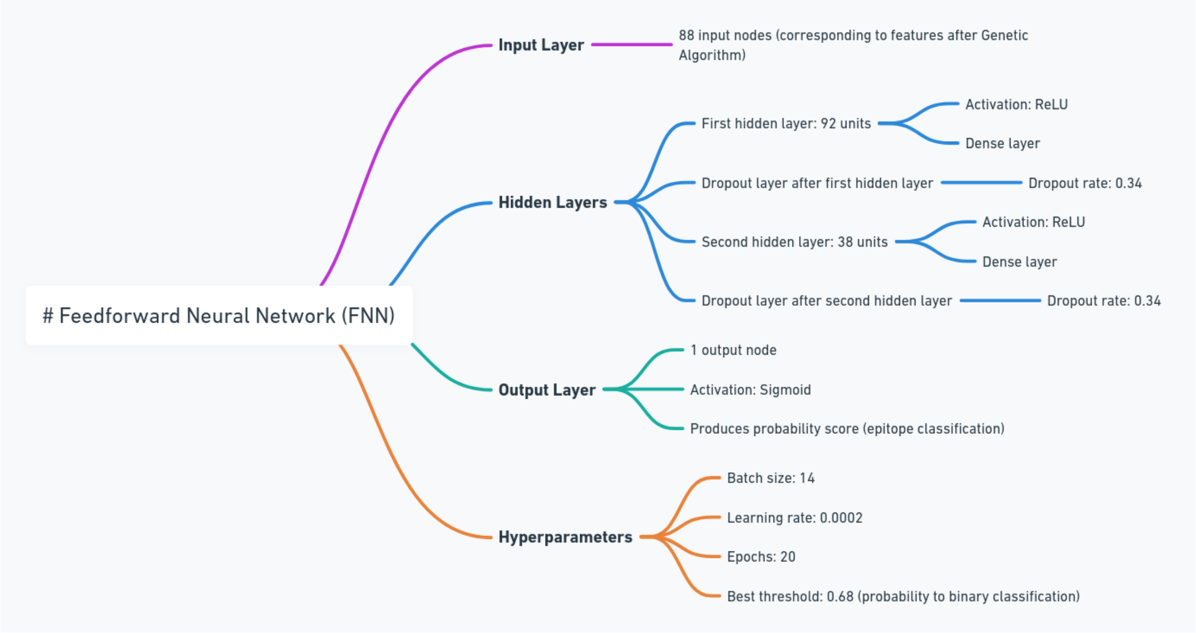
# Feedforward Neural Network (FNN)
Input Layer
88 input nodes (corresponding to features after Genetic Algorithm)
Hidden Layers
First hidden layer: 92 units
Activation: ReLU
Dense layer
Dropout layer after first hidden layer
Dropout rate: 0.34
Second hidden layer: 38 units
Activation: ReLU
Dense layer
Dropout layer after second hidden layer
Dropout rate: 0.34
Output Layer
1 output node
Activation: Sigmoid
Produces probability score (epitope classification)
Hyperparameters
Batch size: 14
Learning rate: 0.0002
Epochs: 20
Best threshold: 0.68 (probability to binary classification)

Figure 1 - FNN Architecture

XGBoost

XGBoost, or Extreme Gradient Boosting, is a powerful ensemble learning technique based on decision trees. It builds upon traditional boosting algorithms, adding regularization and other enhancements to improve performance.

Why XGBoost?

- **Handling High-Dimensional Data**: XGBoost’s ability to effectively handle large, structured datasets made it an ideal choice for this task. Its decision-tree-based approach is well-suited for capturing feature interactions that are particularly important in biological datasets, where relationships are often hierarchical.

- **Ensemble Learning**: The boosting technique involves training multiple weak learners sequentially, each improving upon the errors of the previous one. This iterative refinement allows XGBoost to excel in terms of prediction accuracy and robustness, especially when dealing with noisy data.
- **Class Weights for Imbalance Handling**: During training, **class weights** were assigned to counter the inherent imbalance in the dataset, thereby encouraging the model to focus more on the underrepresented class (epitopes).

XGBoost Model Architecture

The XGBoost model architecture was optimized using **Bayesian Optimization**, a technique that efficiently explores the hyperparameter space to find the best-performing configuration. Bayesian optimization uses probabilistic models to select hyperparameters that are expected to yield the highest performance based on prior evaluations. This method was chosen for its ability to find the optimal parameters with fewer evaluations compared to methods like grid search or random search.

The following key hyperparameters were optimized for the XGBoost model:

- **colsample_bytree:** 0.8799 — This parameter specifies the fraction of features that are randomly sampled for each tree. A value of 0.8799 means that about 88% of the features are considered in each tree, ensuring that important features are included while reducing overfitting.
- **learning_rate:** 0.1856 — The learning rate controls the step size at each iteration of boosting. A value of 0.1856 is relatively high, allowing the model to learn quickly while still maintaining stability in the learning process.
- **max_depth:** 8 — This parameter defines the maximum depth of the trees. A depth of 8 provides a good balance between capturing complex interactions in the data and avoiding overfitting.
- **n_estimators:** 85 — The number of boosting rounds (or trees) in the ensemble. With 85 trees, the model has sufficient capacity to iteratively refine its predictions without being too large, which would increase the risk of overfitting.
- **subsample:** 0.9626 — This parameter controls the fraction of the training data that is randomly sampled to grow each tree. A subsample of 96.26% helps in preventing overfitting by ensuring that each tree is trained on a slightly different subset of the data.

Handling Class Imbalance: SMOTE, Focal Loss, and Class Weights

Imbalanced data poses a major challenge in epitope prediction, as models are prone to becoming biased towards the majority class, leading to poor detection of epitopes.

SMOTE

The **Synthetic Minority Over-sampling Technique (SMOTE)** was applied during data preprocessing to generate synthetic samples of the minority class. SMOTE works by identifying nearest neighbors in the minority class and generating interpolated samples, rather than simply replicating existing instances. This strategy helps prevent overfitting, which often occurs when identical samples are repeated, and ensures that models like XGBoost can develop more generalized decision boundaries.

Focal Loss in Feedforward Neural Network (FNN)

In the training of the Feedforward Neural Network (FNN) model, **Focal Loss** was employed to directly mitigate the issue of class imbalance. Focal Loss, introduced by Lin et al. (2017), extends the traditional Cross-Entropy Loss by reducing the relative loss for well-classified examples (where the predicted probability is close to the true label). This focuses the model's learning on harder-to-classify examples, which often belong to the minority class, such as epitopes in our case.

The **Focal Loss** function is given by the formula:

$$FL(pt) = -\alpha t(1 - pt)\gamma \;\; log(pt)$$

Where:

- pt is the model's predicted probability for the true class label.
- αt is a balancing factor to adjust the importance of the class, typically used to handle class imbalance.
- γ is a modulating factor that adjusts the rate at which easy examples are down-weighted.

When $\gamma = 0$, Focal Loss simplifies to standard Cross-Entropy Loss. As γ increases, the loss for well-classified examples decreases, allowing the model to focus more on misclassified and harder examples.

In this study, Focal Loss was selected for its ability to emphasize learning from the minority class (epitopes), which are crucial to predict accurately in the context of epitope prediction for applications like vaccine development. By reducing the contribution of easy-to-classify non-epitope examples, the model focused on minimizing false negatives, improving recall for the underrepresented epitope class.

The choice of Focal Loss over standard loss functions ensured that the FNN model was more adept at predicting epitopes without being biased towards the majority class, thereby improving its practical utility for immunoinformatics.

Class Weights

In addition to Focal Loss, **Class Weights** were used in both the FNN and XGBoost models to enhance the learning of the minority class. By assigning greater importance to the minority class samples in the loss calculation, the models were encouraged to minimize errors for epitopes more severely than for non-epitopes.

This weighting was integrated into the **loss function** of the FNN and in the **objective function** for XGBoost, effectively altering the optimization process to penalize misclassification of epitopes more heavily. By amplifying the learning impact of underrepresented samples, this approach worked synergistically with Focal Loss and SMOTE to improve the sensitivity of the models.

These combined strategies (SMOTE, Focal Loss, and Class Weights) created a balanced approach to tackle class imbalance. They ensured that the models could recognize patterns in the minority class while maintaining generalization capabilities across the entire dataset, making the predictions both accurate and reliable.

Hyperparameter Tuning: Bayesian Optimization

Once the models were defined, it was crucial to tune their hyperparameters to achieve optimal performance. **Bayesian Optimization** was utilized for this purpose, leveraging its probabilistic approach to systematically search for the best combination of hyperparameters for each model.

Why Bayesian Optimization?

Hyperparameter tuning is essential to improve model accuracy, avoid overfitting, and ensure the robustness of the models. Traditional tuning techniques, such as **grid search** or **random search**, are often inefficient due to their exhaustive or random nature, especially when dealing with numerous hyperparameters.

Bayesian Optimization addresses these challenges through a more strategic approach. It builds a probabilistic model (typically a Gaussian Process) of the objective function and uses it to determine which hyperparameters to evaluate next, based on expected improvement. This is particularly advantageous for models like FNN and XGBoost, where:

- The hyperparameter space is vast (e.g., number of layers, learning rate, activation functions for FNN, and tree depth, learning rate, number of estimators for XGBoost).
- Evaluations are computationally expensive.

Bayesian Optimization Process

1. **Surrogate Model Construction**: Bayesian Optimization first constructs a **surrogate model**—an approximation of the objective function based on previous evaluations. This model

is used to predict the performance of different hyperparameter combinations without explicitly training the model.

2. **Acquisition Function**: The next hyperparameters to evaluate are selected based on an **acquisition function**, which quantifies the expected improvement. This balances **exploration** (searching new areas of the hyperparameter space) and **exploitation** (refining areas known to yield good results).
3. **Hyperparameter Search**: Bayesian Optimization was used to tune the number of **hidden layers**, **neurons per layer**, **dropout rates**, and **learning rates** in the FNN, as well as **tree depth**, **learning rate**, **subsampling ratios**, and **colsample_bytree** for XGBoost. This efficient search led to a significant reduction in computational cost compared to exhaustive methods.

The implementation of Bayesian Optimization resulted in a systematic exploration of hyperparameter spaces, leading to models that were both computationally efficient and high performing. For example, in FNN, an optimal configuration of neurons and layers could achieve a balance between complexity and generalizability, while in XGBoost, parameters such as maximum depth and learning rate were fine-tuned to control model complexity and prevent overfitting.

Model Evaluation and Metrics

Model evaluation was conducted using a suite of metrics designed to provide a comprehensive understanding of model performance, especially considering the imbalanced nature of the dataset. The evaluation focused on metrics that could capture both the model's overall accuracy and its effectiveness in predicting the minority class (epitopes).

Mathews Correlation Coefficient (MCC)

Matthews Correlation Coefficient (MCC) is an effective measure for assessing model performance in imbalanced datasets. Unlike accuracy, which can be misleading when the majority class dominates, MCC provides a balanced evaluation that takes into account true positives, true negatives, false positives, and false negatives.

MCC values range from -1 (complete misclassification) to +1 (perfect classification), with 0 indicating a prediction no better than random. For this research, MCC was used as a key performance metric due to its ability to provide an unbiased view of how well the models predict both classes, making it particularly suitable for the epitope prediction task.

Precision, Recall, and F1 Score

- **Precision**: Precision was calculated to determine how many of the instances predicted as epitopes were correctly identified. This metric is crucial for minimizing **false positives**, which is especially important when predicting epitopes to avoid unnecessary downstream experimental validation.

- **Recall (Sensitivity)**: Recall was used to evaluate how well the model identified all true epitopes, thus minimizing **false negatives**. A high recall is critical in biological research to ensure that all possible epitopes are included, thereby avoiding the exclusion of potential vaccine targets.
- **F1 Score**: The **F1 Score**, which is the harmonic mean of Precision and Recall, was employed to provide a balanced view of the model's performance. This metric is particularly useful when the goal is to find an optimal balance between identifying as many true positives as possible without introducing too many false positives.

Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

The **AUC-ROC** was used to evaluate the discriminative ability of the models. The **ROC curve** plots the true positive rate (recall) against the false positive rate, providing a visual representation of the trade-off between sensitivity and specificity across various threshold levels. The **Area Under the Curve (AUC)** gives a single metric summarizing the model's capability to differentiate between epitopes and non-epitopes. A high AUC value indicates that the model has a good balance between correctly identifying positive cases and minimizing false positives, which is crucial for epitope identification.

Project Management

The successful execution of this research required careful project management, primarily concerning the computational challenges and resource planning due to the high computational demands of the models and dataset. This section outlines the project management strategies employed, with a focus on managing computational resources and leveraging cloud computing to ensure the timely and efficient completion of experiments.

Computational Challenges and Cloud Computing Considerations

One of the major project management aspects involved addressing the significant computational requirements inherent to the research. Given the complexity of the dataset, the need for multiple rounds of hyperparameter tuning, and the development of ensemble and deep learning models, traditional on-premises infrastructure would have posed considerable limitations.

To overcome these challenges, cloud computing resources were utilized, specifically leveraging Azure Machine Learning (Azure ML) services for managing, training, and optimizing models. This ensured that the project adhered to timelines and quality standards by optimizing the computing environment effectively.

Azure Machine Learning Workspace

Azure Machine Learning Workspace played a pivotal role in managing various phases of the machine learning workflow, providing several key benefits that facilitated efficient project management:

- Scalability: Azure ML's elastic cloud infrastructure enabled scalable compute resources, which were crucial for training deep learning models like Feedforward Neural Networks (FNN), which require substantial GPU power. The ability to dynamically adjust compute resources based on workload ensured that computational bottlenecks did not disrupt the project timeline.
- Experiment Management: The built-in experiment tracking capabilities of Azure ML enabled the management of multiple experiments involving different hyperparameter configurations and feature subsets. This allowed for streamlined coordination of Bayesian Optimization across numerous model configurations and effectively minimized redundant computation, thus managing time and resources efficiently.
- Data Accessibility and Collaboration: Azure ML provided centralized data management, where the dataset was stored in tabular format, ensuring consistency across experiments. The accessibility of this dataset for both local and remote processing facilitated collaborative efforts, which is crucial in a complex project where multiple stages of analysis and development occur concurrently.

Addressing Computational Challenges

The following strategies were employed to address computational challenges, showcasing the role of effective project management in ensuring that resource constraints did not impede progress:

- Training Complexity: The training of both the FNN and XGBoost models, especially with Bayesian Optimization for hyperparameter tuning, involved a large number of iterations. On-premises hardware would not have been feasible to handle these demands efficiently. Using Azure ML's compute clusters significantly reduced training times and allowed for the evaluation of multiple hyperparameter combinations within practical limits, thereby adhering to project timelines.
- Parallel Execution: Azure ML's parallelism capabilities were leveraged to execute multiple instances of Genetic Algorithm (GA)-based feature selection and model training concurrently. This not only saved time but also ensured a thorough exploration of feature and hyperparameter spaces, leading to a more optimal model development process. This approach ensured that each stage of the experiment pipeline progressed without unnecessary delays, thus maintaining the overall schedule.

The strategic use of cloud computing and project management tools such as Azure ML was instrumental in overcoming computational limitations and ensuring that the entire pipeline— from data preprocessing to model evaluation—was executed under optimal conditions. The flexibility and computational power provided by Azure ML allowed the research to meet its objectives within the desired timeframe, despite the computational challenges and intensive resource requirements.

Results and Discussion

This section presents a detailed discussion of the results obtained from training two different models—the Feedforward Neural Network (FNN) and XGBoost—on the coronavirus specific subset. Additionally, it evaluates the generalization of the XGBoost model on the heterogeneous dataset to assess the model's broader applicability. The focus is on understanding how feature selection, model training, and specific configurations influenced model performance.

Results

Feature Selection Results

Feature selection was a critical part of the pipeline, aimed at reducing the dimensionality of the dataset while retaining the most informative features to improve model performance.

- Initially, the complete dataset included 393 features, encompassing various biochemical properties of amino acids. The Boruta algorithm was employed first to eliminate irrelevant or redundant features, narrowing the set to 195 features, which represented a 49.74% reduction in feature size..
- Subsequently, a **Genetic Algorithm (GA)** was used to further refine the set, reducing the number to 88 features for the coronavirus subset. This additional feature reduction resulted in a 77.32% reduction from the original set, allowing the models to focus on the most impactful features.
- **Figure 2** (Feature Correlation Heatmap After Boruta) demonstrates the impact of Boruta's filtering on feature correlation. Compared to the initial matrix, the number of highly correlated features is visibly reduced, which ensures a more interpretable and less redundant feature set. This improved interpretability aids the subsequent model training process by simplifying the data while retaining the critical information needed for prediction.

- The GA was further analyzed to evaluate performance over successive generations. **Figure 3** (GA Accuracy Over Generations with Early Stopping) depicts how the GA- based feature selection optimized the feature set across generations. Early stopping was applied to prevent overfitting, and the accuracy peaked at Generation 2, highlighting the efficacy of GA in identifying optimal feature combinations within a few iterations.

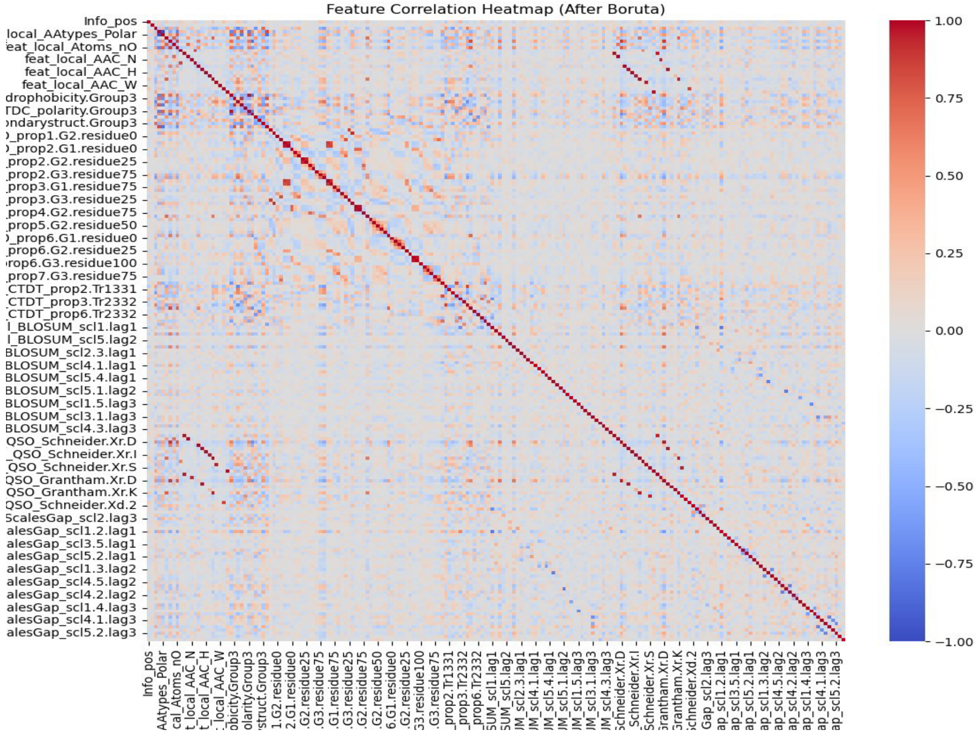

Figure 2 – Feature Correlation Heatmap after Boruta

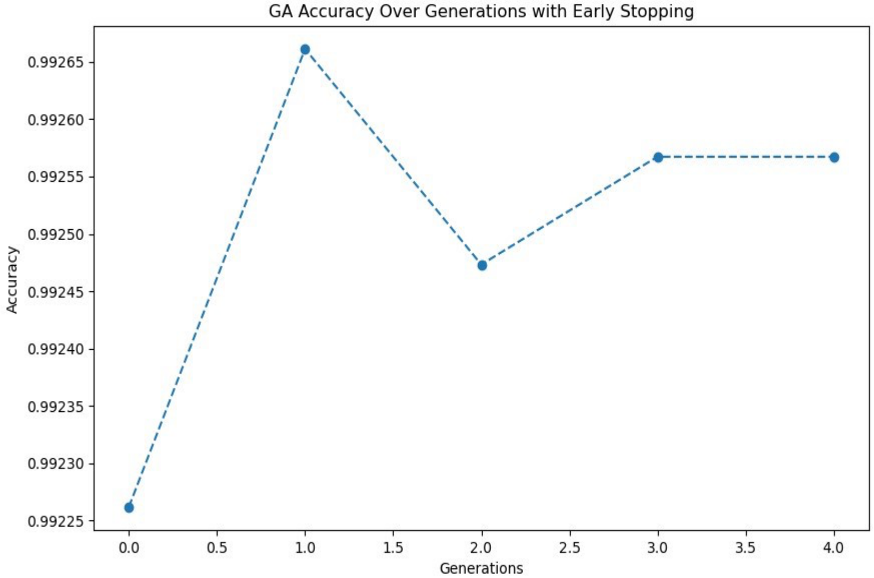

Figure 3 – GA accuracy over generation with early stopping

Model Performance on the Target Organism (Coronavirus)

XGBoost Performance on Coronavirus Subset

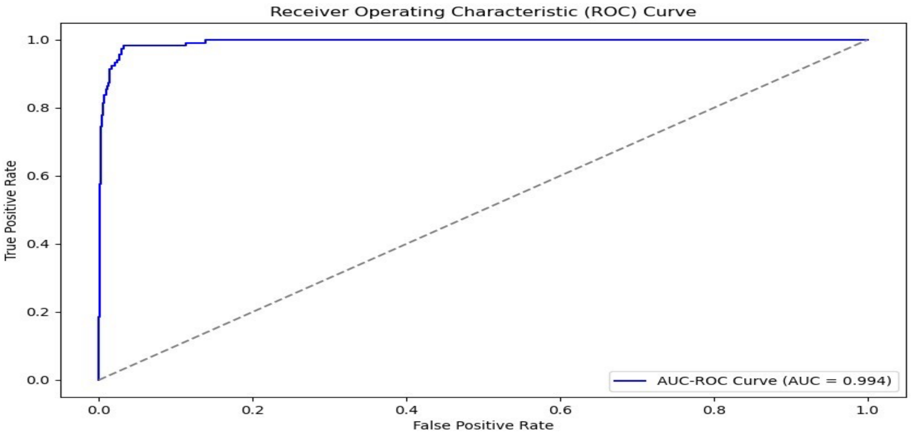

Figure 4 - ROC curve for XGBoost on Target Organism

The XGBoost model demonstrated excellent performance on the coronavirus dataset. The Receiver Operating Characteristic (ROC) Curve for XGBoost is shown in **Figure 4**, which indicates an **AUC of 0.994**. This near-perfect score reflects the model's high ability to discriminate between epitopes and non-epitopes. The ROC curve, hugging the top-left corner, confirms the model's robustness in classification.

- **Confusion Matrix Analysis**:

  - **Figure 5** provides the confusion matrix for XGBoost on the coronavirus specific dataset. The model correctly classified **5282 negative instances** and **94 positive instances**, with only **25 false positives** and **24 false negatives**. This yields an **accuracy of 99%**, an **F1 score of 0.88**, and a **Matthews Correlation Coefficient (MCC) of 0.789**.
  - The low false positive and false negative rates indicate effective learning and an ability to distinguish well between the positive and negative classes. This is further supported by the **Precision-Recall Curve** shown in **Figure 6**, where a high precision level is maintained across different recall thresholds, signifying the model's capability to keep false positives at a minimum while ensuring a high true positive rate.

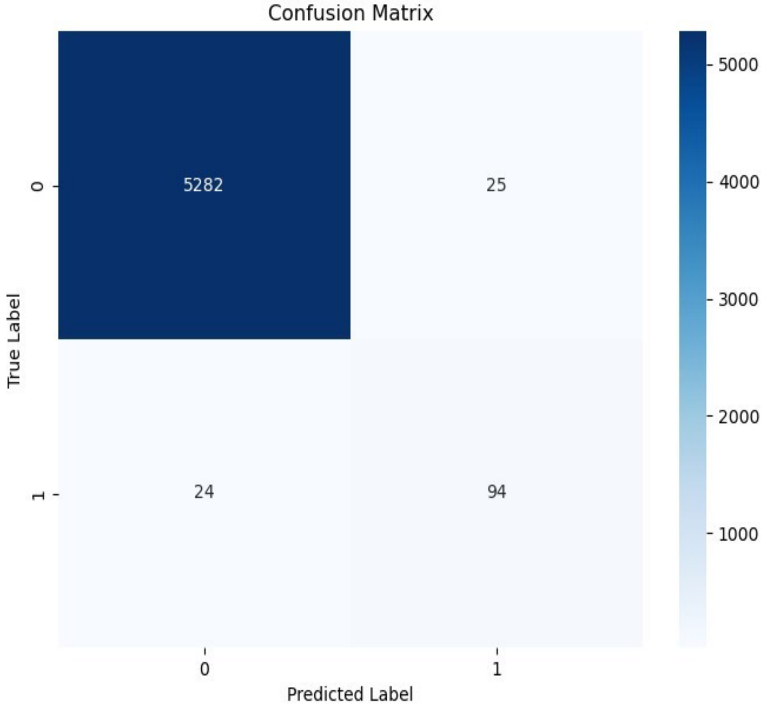

Figure 5- Confusion Matrix for XGBoost on Target Organism

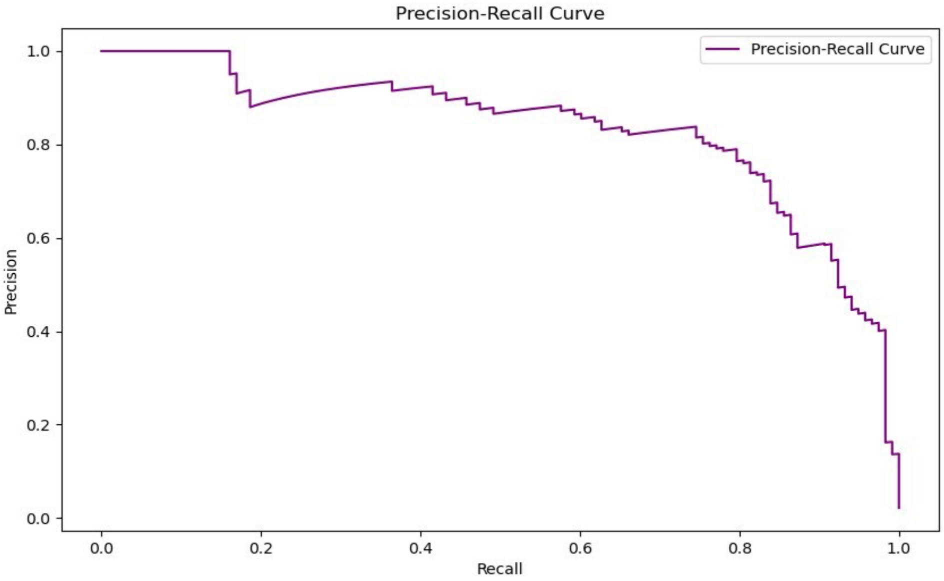

Figure 6- Precision-Recall Curve for XGBoost on Target Organism

FNN Performance on Coronavirus Subset

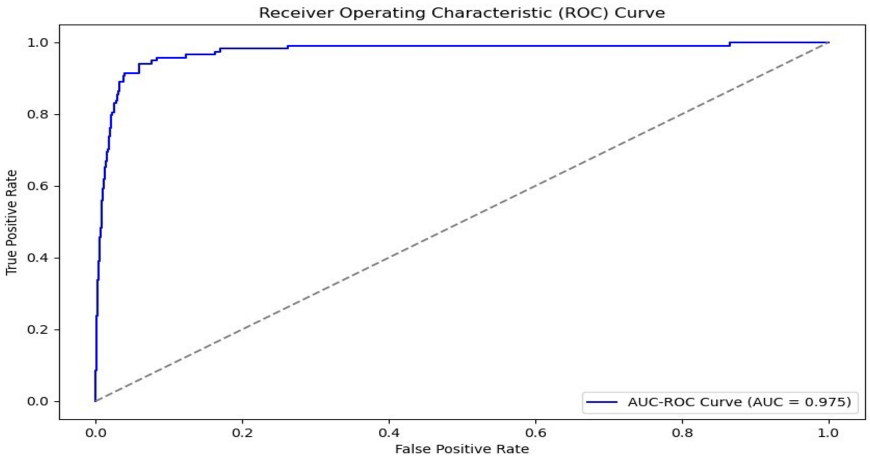

Figure 7 - ROC Curve for FNN on Target Organism

The Feedforward Neural Network (FNN) showed relatively lower performance compared to XGBoost. **Figure 7** (ROC Curve for FNN) shows an **AUC of 0.975**, which, while still strong, lags behind XGBoost. The ROC curve reflects the FNN's limitations in achieving the same level of discriminative power as XGBoost, particularly in identifying borderline cases.

- **Confusion Matrix Analysis**:
  - **Figure 8** presents the confusion matrix for the FNN model. It correctly classified **5257 negative instances** but had **50 false positives**. In the case of the positive instances, **70 true positives** were identified, while **48 were misclassified**. The FNN model achieved an **F1 score of 0.71**, an **accuracy of 98.7%**, and an **MCC of 0.579**, notably lower than XGBoost, indicating less reliable predictive capability.
  - The **Precision-Recall Curve** in **Figure 9** reveals a lower precision as recall increases, suggesting that FNN struggled to maintain a balance between correctly identifying epitopes and avoiding false positives. The drop in precision as recall rises highlights the FNN's reduced sensitivity when handling complex feature interactions compared to the XGBoost model.

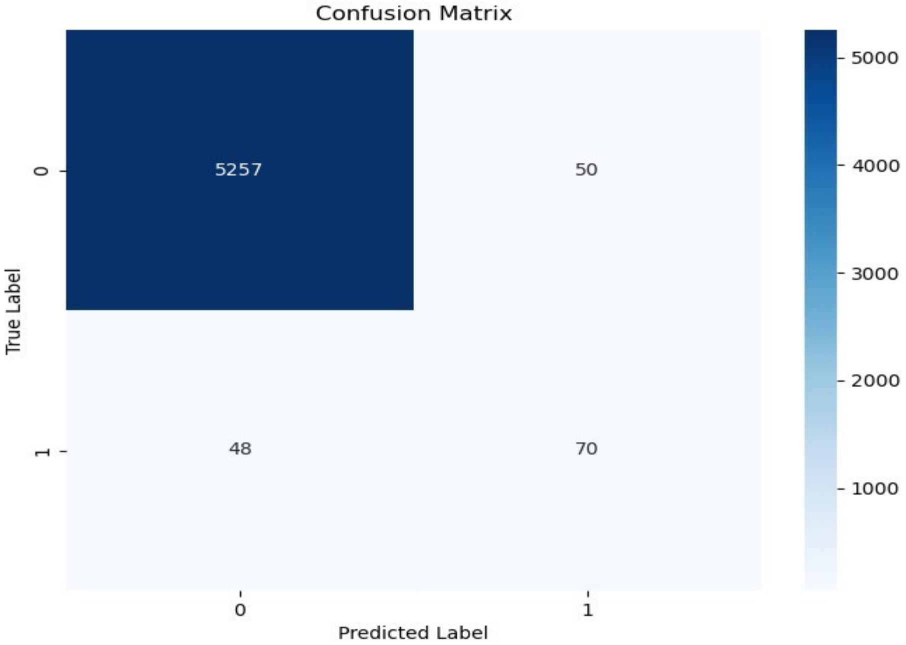

Figure 8 - Confusion Matrix for FNN on Target Organism

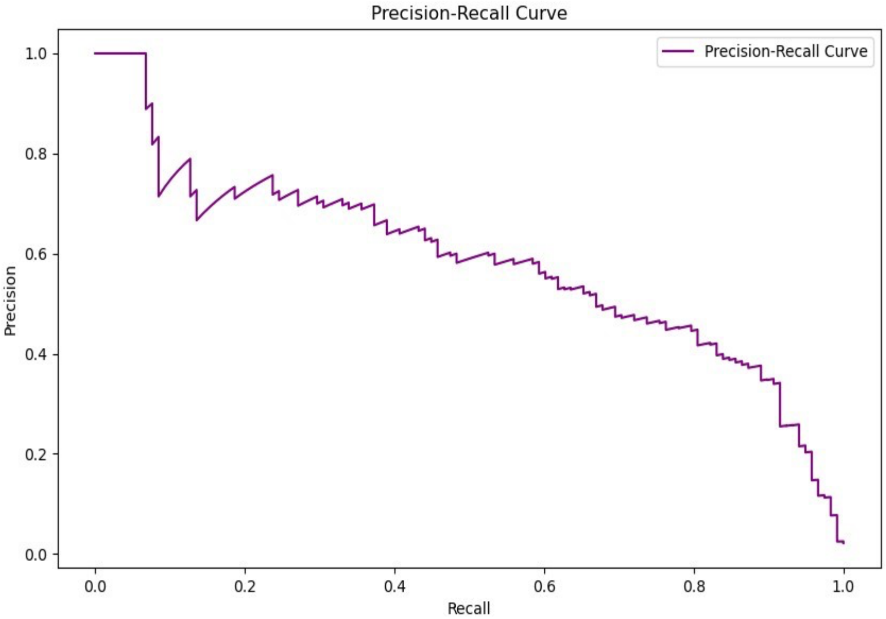

Figure 9 - Precision-Recall Curve for FNN on Target Organism

Model Generalization on the Heterogeneous Dataset

The XGBoost model, having demonstrated superior performance on the coronavirus-specific dataset, was then tested for its ability to generalize across a heterogeneous dataset containing multiple pathogens.

XGBoost Generalization Performance

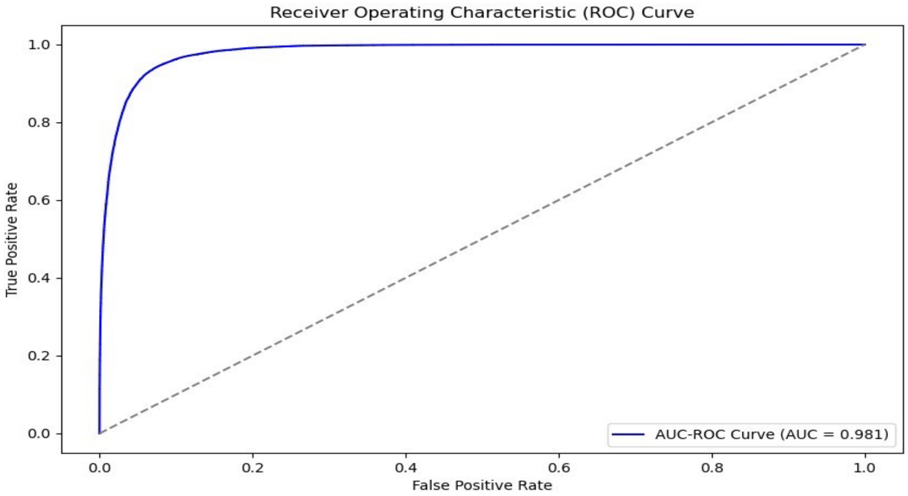

Figure 10 – ROC Curve for XGBoost on Heterogenous Data

The generalization results are illustrated in **Figure 10** (**ROC Curve for XGBoost on Heterogeneous Data**), which shows an **AUC of 0.981**. This high AUC indicates that the model retained strong discriminative power even when exposed to data from diverse organisms, underscoring its generalizability.

- **Confusion Matrix Analysis**:
  - **Figure 11** presents the confusion matrix for the heterogeneous dataset. The model correctly classified **90,329 negative instances** and **22,850 positive instances**. There were **4,112 false positives** and **2,948 false negatives**, resulting in an **accuracy of 94%**, an **F1 score of 0.87**, and an **MCC of 0.829**.
  - Although the number of misclassifications increased compared to the coronavirus-specific training, the relatively low proportion of false positives and negatives highlights that the model successfully generalized to other pathogens. The model's performance indicates that the features selected during training effectively captured information relevant across various organisms.
  -

- **Precision-Recall Analysis**:
  - **Figure 12** (**Precision-Recall Curve for XGBoost on Heterogeneous Data**) shows that precision was maintained at a high level across varying recall rates, which is promising for practical applications. The curve's shape implies that XGBoost could reliably identify true epitopes even when exposed to more complex, multi-pathogen data.

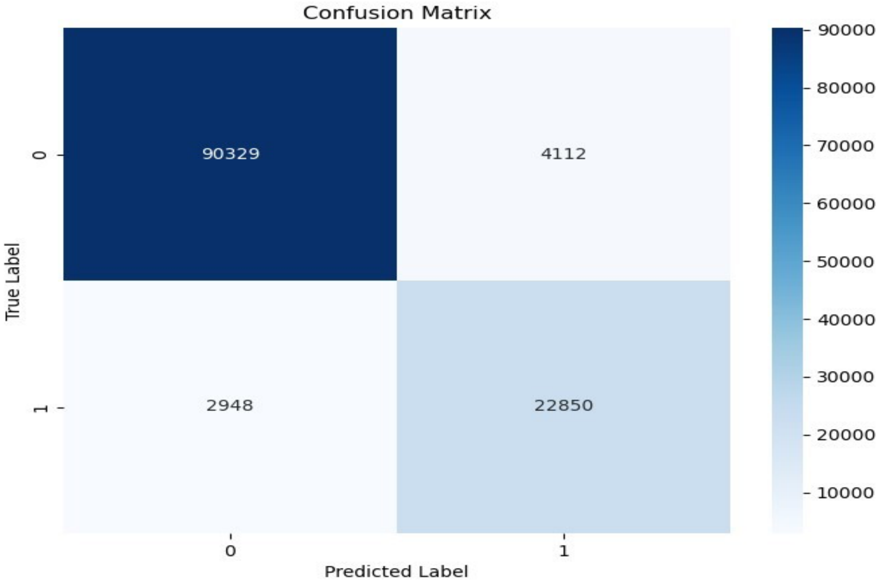

Figure 11 – Confusion Matrix for XGBoost on Heterogenous Data

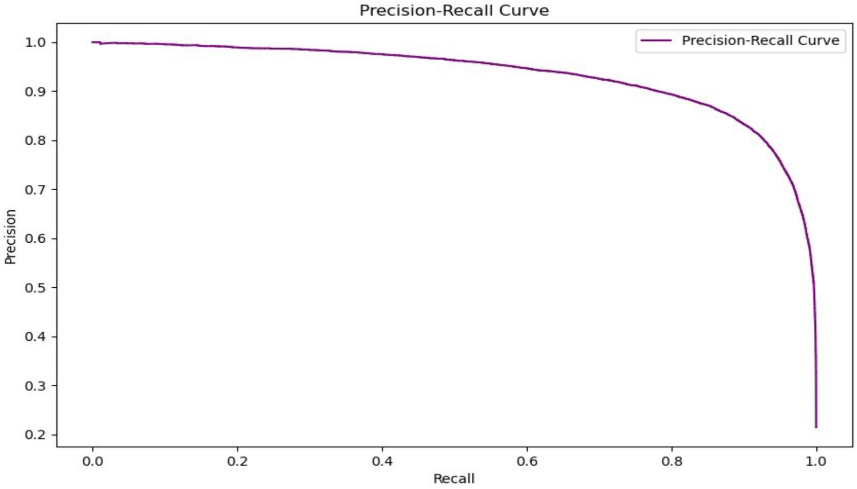

Figure 12 – Precision- Recall Curve for XGBoost on Heterogenous Data

Discussion

Impact of Feature Selection

The two-step feature selection process—Boruta followed by Genetic Algorithm (GA)— proved crucial in reducing dimensionality while preserving predictive power, which directly impacted the success of the XGBoost model.

- **Boruta's Contribution**: As observed from **Figure 2** (Feature Correlation Heatmap After Boruta), Boruta effectively reduced redundancy by eliminating highly correlated features, which, if retained, could lead to overfitting. By focusing on the most relevant features, Boruta enabled both models to operate with a more streamlined dataset.
- **GA Optimization**: The GA accuracy curve (**Figure 3**) shows how the GA continued to optimize the feature set over generations, peaking by **Generation 2**. This emphasizes the importance of an evolutionary approach in searching for optimal feature combinations, allowing the models to achieve enhanced accuracy with fewer features. The use of early stopping ensured that overfitting was minimized, making the feature set more robust for both training and generalization.

Comparison of Model Performance on the Target Organism

- **XGBoost vs. FNN**:
  - The XGBoost model outperformed FNN in nearly all key metrics, as demonstrated by the ROC curves in **Figures 4 and 7**, and the confusion matrices in **Figures 5 and 8**. XGBoost's higher **AUC (0.994)** compared to FNN's **AUC (0.975)** indicates a better ability to separate positive and negative classes. The **MCC values** (**0.789 for XGBoost vs. 0.579 for FNN**) further confirm this finding, suggesting that XGBoost was more balanced in its classification decisions.
  - The **precision-recall comparison** is also telling. The **Precision-Recall Curve for XGBoost** (**Figure 6**) remains high across recall thresholds, while **Figure 10** for FNN shows a marked decline, indicating that XGBoost maintained more reliable predictions without compromising sensitivity.

Generalizability of the Best Model

After demonstrating superior performance on the coronavirus-specific subset, the XGBoost model was assessed for generalizability across a more heterogeneous dataset.

- **Performance Across Pathogens**:
    - The results in **Figure 10** (ROC Curve for Heterogeneous Dataset) and **Figure 11** (Confusion Matrix) indicate that XGBoost was successful in generalizing to a broader dataset, with an **AUC of 0.981** and an **MCC of 0.829**. These metrics suggest that the model retained a significant portion of its predictive capability despite the increased complexity of the data.
- **Precision-Recall Considerations**:
    - The **Precision-Recall Curve** (**Figure 12**) for the heterogeneous dataset reveals that XGBoost maintained a high level of precision across a range of recall thresholds, indicating that the model was able to correctly identify true epitopes while maintaining a relatively low rate of false positives. The shape of the precision-recall curve demonstrates a robust trade-off, ensuring that the model remains reliable for epitope detection even when applied to a diverse dataset.
- **Effectiveness of Feature Selection for Generalizability**:
    - One critical factor contributing to the generalization success of XGBoost was the stratified sampling used for Boruta feature selection. By using a 20% representative sample of the heterogeneous dataset, the model was trained on features that effectively captured the variability present in different pathogens. This stratified approach ensured that the selected features were reflective of the entire dataset's diversity without imposing prohibitive computational costs.
- **Addressing Computational Challenges**:
    - **Batch Processing and Undersampling**: To mitigate computational challenges, batch processing was used for training the XGBoost model on the heterogeneous dataset. This approach allowed the training process to be broken into manageable chunks, thereby reducing memory usage and allowing the model to be trained on available computational infrastructure. This strategy was crucial in maintaining the feasibility of training without compromising model performance.
    - Furthermore, majority undersampling was employed to manage class imbalance, as oversampling would have significantly increased the dataset size and computational cost. Although undersampling might have led to a loss of information from the majority class, the model's high **MCC (0.829)** and **F1 score (0.87)** indicate that the approach was effective in maintaining a balance between precision and recall for minority classes.

Practical Implications and Contributions

The findings of this study have significant implications for the field of immunoinformatics and vaccine development:

- **Organism-Specific Models with Generalizability**:
  - The results challenge the common belief that organism-specific models are inherently limited in their ability to generalize to other datasets. The XGBoost model, trained specifically on coronavirus epitopes, successfully generalized to a dataset containing a variety of pathogens, suggesting that well-optimized models can extend their utility beyond their initial scope.
  - This finding implies that machine learning models developed for specific pathogens can serve as foundational models that can be adapted to other similar pathogens, thus providing a faster and more cost-effective way to respond to emerging infectious diseases.
- **Optimization Techniques for Real-World Applicability**:
  - The combination of Boruta and GA for feature selection has proven to be both effective and computationally feasible. The Boruta-GA combination enabled the identification of a compact yet informative set of features, significantly reducing dimensionality while maintaining model accuracy. This reduction is crucial for training models in a practical timeframe, especially when dealing with high-dimensional biological datasets.
  - By implementing strategies like batch processing and undersampling, this study also demonstrates how computational challenges can be addressed in real-world settings where access to high-performance computing may be limited. These approaches provide a blueprint for researchers aiming to optimize models in constrained computational environments.
- **Implications for Vaccine and Therapeutic Development**:
  - The ability of the XGBoost model to generalize across different pathogens means that predictions regarding B-cell epitopes can be extended beyond a single organism. This capability can be instrumental in vaccine development, where identifying conserved epitopes that elicit immune responses across multiple virus strains is often the goal.
  - Furthermore, the study's approach to feature selection and model optimization can guide the development of predictive models for other pathogens. By focusing on relevant biological features and using stratified sampling, similar models can be designed for other emerging pathogens, thereby accelerating the discovery of candidate epitopes for vaccine and therapeutic antibody production.

Conclusion and Recommendation

Conclusion

The primary objective of this study was to investigate the **extent to which organism-specific training** for predicting linear B-cell epitopes (LBCEs), particularly for Coronavirus, can improve model performance. This research focused not only on evaluating whether pathogen-specific models outperform generalist models but also on quantifying **how much better** the pathogen-specific models perform. At the same time, the study sought to develop a generalized approach that could maintain high performance across various pathogens. Using advanced machine learning techniques, such as **feature selection** via Boruta and **Genetic Algorithms (GA)**, the study optimized models for both organism-specific accuracy and cross-pathogen applicability. These results offer significant insights into computational epitope prediction, vaccine development, and therapeutic antibody design.

Organism-Specific vs. Generalist Models

The results demonstrate that organism-specific models, particularly those trained on the Coronavirus subset, **significantly outperformed** generalist models across multiple metrics. As shown in **Figure 13**, the XGBoost model achieved an **AUC of 0.994**, indicating near-perfect discriminative ability between epitopes and non-epitopes. Additionally, the model's **F1 score** of **0.88** and **Matthews Correlation Coefficient (MCC)** of **0.789** highlight its **superior precision** in classification tasks compared to the generalist models. This detailed comparison (Figure 12) provides a clear illustration of **how much better** the organism-specific model performed, rather than just whether it performed better.

In contrast, the Feedforward Neural Network (FNN), while still effective, exhibited a notable gap in performance compared to XGBoost. The FNN recorded an **AUC of 0.975** and an **F1 score of 0.71**, underscoring its relative limitations in precision, recall, and generalizability. These findings suggest that although neural networks are powerful for modeling non-linear relationships, XGBoost's ensemble approach, particularly its use of decision trees and boosting, was more adept at handling the biological complexity and feature interactions within the dataset.

Generalization to Heterogeneous Data

The generalizability of the models was tested on a heterogeneous dataset containing multiple pathogens, including Flu, Epstein-Barr virus, and Lentivirus. Again, **Figure 13** highlights that the XGBoost model continued to demonstrate its **superiority**, achieving an **AUC of 0.981**, an **F1 score of 0.87**, and an **MCC of 0.829**. These results confirm that XGBoost retained a significant portion of its predictive power even when exposed to organisms outside its training set. This crucial finding suggests that **organism-specific models**, when optimized, can generalize effectively to other pathogens, broadening their applicability beyond the target organism.

The results from this generalization study underscore the importance of **feature selection** and **model optimization**. The stratified sampling approach used during Boruta feature selection was essential in ensuring that the model captured diverse, relevant features across different pathogens, thus contributing to its robust performance in the heterogeneous dataset.

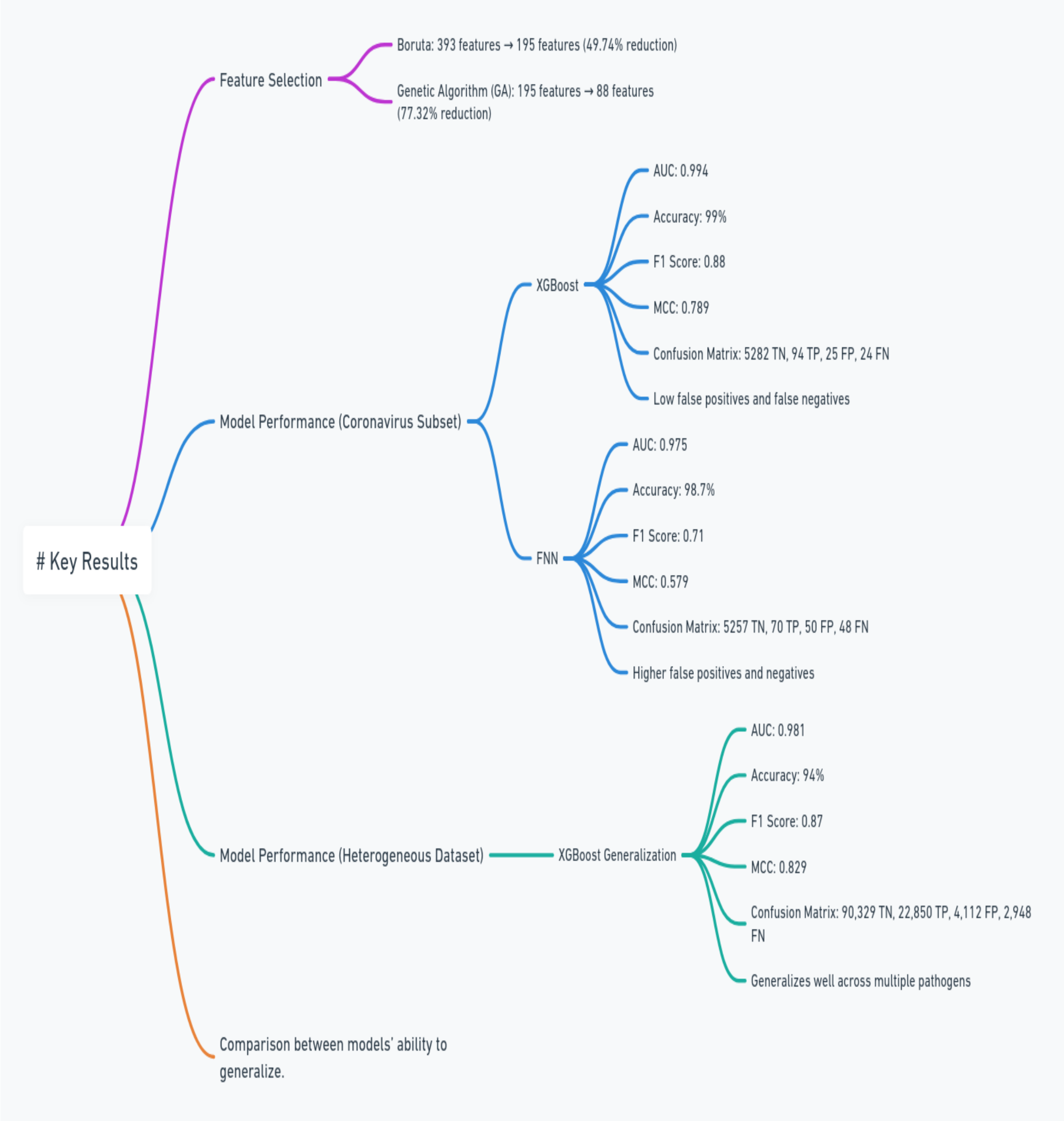
# Key Results
Feature Selection
Boruta: 393 features → 195 features (49.74% reduction)
Genetic Algorithm (GA): 195 features → 88 features (77.32% reduction)
Model Performance (Coronavirus Subset)
XGBoost
AUC: 0.994
Accuracy: 99%
F1 Score: 0.88
MCC: 0.789
Confusion Matrix: 5282 TN, 94 TP, 25 FP, 24 FN
Low false positives and false negatives
FNN
AUC: 0.975
Accuracy: 98.7%
F1 Score: 0.71
MCC: 0.579
Confusion Matrix: 5257 TN, 70 TP, 50 FP, 48 FN
Higher false positives and negatives
Model Performance (Heterogeneous Dataset)
XGBoost Generalization
AUC: 0.981
Accuracy: 94%
F1 Score: 0.87
MCC: 0.829
Confusion Matrix: 90,329 TN, 22,850 TP, 4,112 FP, 2,948 FN
Generalizes well across multiple pathogens
Comparison between models' ability to generalize.

Figure 13- Pipelines Performance Comparison

Impact of Feature Selection

The two-step feature selection process, combining Boruta and GA, proved to be a highly effective strategy for reducing the dimensionality of the dataset while preserving predictive power. Boruta's elimination of irrelevant and redundant features was instrumental in reducing noise and potential overfitting, resulting in a more manageable and interpretable feature set for model training. This was further enhanced by GA's optimization, which identified the most impactful feature combinations, ensuring that the final models were not only accurate but also generalizable.

The GA accuracy curve showed that feature selection was optimized within a few iterations, peaking at Generation 2, which minimized the risk of overfitting while maintaining robust model performance. The use of early stopping during GA iterations prevented the model from becoming overly complex, further contributing to its generalizability across different datasets.

Practical Implications

This research has significant practical implications, particularly in the fields of immunoinformatics and vaccine development. By demonstrating that organism-specific models can generalize to other pathogens, this study challenges the common assumption that such models are limited in scope. This finding is particularly relevant in the context of emerging infectious diseases, where rapid adaptation to new pathogens is crucial.

The ability of the XGBoost model to predict LBCEs across multiple pathogens implies that well-optimized machine learning models can serve as foundational tools for broader applications in vaccine and therapeutic development. For instance, predictive models trained on specific pathogens like Coronavirus can be adapted to other similar pathogens, offering a faster and more cost-effective approach to vaccine design.

Moreover, the success of feature selection techniques, such as Boruta and GA, in this study provides a blueprint for optimizing feature sets in other biological prediction tasks. By reducing dataset complexity while retaining critical information, these techniques enable the development of more efficient and accurate predictive models, even in computationally constrained environments.

Recommendations

Based on the findings of this research, several recommendations can be made to further improve the predictive power of LBCE models and enhance their applicability in real-world scenarios.

Leveraging Hybrid and Ensemble Models for Improved Generalizability

While XGBoost performed exceptionally well in both organism-specific and heterogeneous datasets, there is potential to enhance model performance further by integrating hybrid or ensemble approaches. For example, combining the strengths of XGBoost with deep learning models like convolutional neural networks (CNNs) or long short-term memory (LSTM) networks could provide more robust predictions by capturing both local and global dependencies within the protein sequences.

Ensemble learning, which involves combining multiple predictive models to reduce overfitting and improve generalization, should be explored in future studies. Hybrid models that integrate both sequence-based and structure-based features could offer a more comprehensive predictive framework, ensuring that all relevant biological features are considered during epitope prediction.

Addressing Class Imbalance Through Advanced Techniques

Class imbalance remains a persistent challenge in epitope prediction, as evidenced by the need for techniques like Synthetic Minority Over-sampling Technique (SMOTE), Focal Loss, and Class Weights to ensure that the models could effectively learn from the minority class (epitopes). While these methods were successful in mitigating the effects of imbalance, future research could explore more advanced techniques, such as adaptive synthetic sampling (ADASYN) or ensemble methods specifically designed for imbalanced data.

Additionally, using a more sophisticated loss function, such as a weighted F1 score, could further improve the sensitivity of models to epitopes without sacrificing precision. This would be particularly beneficial in applications where identifying all possible epitopes is critical, such as in vaccine development or therapeutic antibody design.

Exploring Transfer Learning for Pathogen Adaptation

One of the key challenges in epitope prediction is the rapid adaptation of models to new and emerging pathogens. Transfer learning, where a model trained on one pathogen is fine-tuned to adapt to another, offers a promising solution to this issue. By leveraging pre-trained embeddings from models like ProtBERT or ProtTrans, future studies could explore how these models can be adapted to novel pathogens with minimal retraining.

Transfer learning has already shown success in related fields of computational biology, and its application in epitope prediction could significantly reduce the time and computational resources required to develop models for new pathogens. This approach would be particularly valuable in pandemic situations, where time is of the essence in developing vaccines and therapies.

Incorporating Structural Data for Conformational Epitope Prediction

While this study focused on linear B-cell epitopes, the prediction of conformational epitopes remains a significant challenge due to the reliance on three-dimensional structural data. Future research should explore the integration of structure-based approaches, such as homology modeling or Cryo-EM data, into machine learning models to predict conformational epitopes.

Tools like ElliPro and Discotope, which rely on protein structural data, could be combined with sequence-based methods to develop hybrid models capable of predicting both linear and conformational epitopes. By incorporating structural features into the predictive models, researchers can expand the scope of epitope prediction beyond linear sequences, thereby increasing the utility of these models in vaccine design and therapeutic development.

Enhancing Computational Efficiency Through Cloud Computing

This study made extensive use of cloud computing resources, particularly Azure Machine Learning, to manage the computational demands of training and optimizing the models. Given the success of this approach, it is recommended that future research projects in computational biology continue to leverage cloud-based platforms to scale their experiments effectively.

Cloud computing offers several advantages, including scalability, flexible resource allocation, and centralized data management, all of which are crucial for handling large biological datasets. By adopting cloud-based infrastructures, researchers can overcome the limitations of on-premises hardware and ensure that their models are trained efficiently, even when dealing with high-dimensional data.

Collaborating Across Disciplines for Broader Impact

The findings of this research underscore the importance of interdisciplinary collaboration in computational biology. By combining expertise from fields such as immunology, bioinformatics, and machine learning, future studies can develop more sophisticated models that address the complex nature of epitope prediction.

Collaborating with experimental biologists, for example, could facilitate the validation of predicted epitopes in laboratory settings, thereby bridging the gap between computational predictions and real-world applications. Additionally, partnerships with pharmaceutical companies and healthcare institutions could accelerate the translation of epitope prediction models into practical tools for vaccine and therapeutic development.

Limitations and Future Directions

While this study offers valuable insights into LBCE prediction, there are several limitations that should be addressed in future research.

Data Quality and Availability

The quality of the training data plays a critical role in the performance of machine learning models. In this study, data were derived from the Immune Epitope Database (IEDB), which, while comprehensive, may contain biases due to variability in experimental conditions or the limited availability of high-quality epitope data for certain pathogens. Future studies could benefit from expanding the dataset to include more diverse and well-validated epitopes, potentially through collaborations with experimental laboratories. Additionally, the use of simulated or synthetic data could supplement real-world datasets, allowing for more extensive training and validation of the models.

Model Interpretability

One of the challenges in using machine learning models, particularly complex models like XGBoost and FNN, is the lack of interpretability. While the models in this study achieved high performance, understanding the biological rationale behind their predictions remains limited. Future research should explore techniques such as SHAP (Shapley Additive explanations) or LIME (Local Interpretable Model-agnostic Explanations) to improve the interpretability of the model's decision-making processes. This would provide more biologically meaningful insights and allow researchers to validate predictions based on biological plausibility.

Expanding to Other Immune Cells

While this study focused on predicting B-cell epitopes, future work could expand to include T-cell epitopes, which are equally critical in the immune response. Incorporating T-cell epitope prediction into the pipeline would provide a more holistic view of the immune response, allowing researchers to identify potential vaccine candidates that elicit both humoral and cellular immunity. Combining B-cell and T-cell epitope prediction models could lead to more comprehensive vaccine designs.

Real-World Applications and Validation

One of the limitations of this study is the lack of experimental validation for the predicted epitopes. Future studies should prioritize collaborations with experimental researchers to test the predicted epitopes in vitro and in vivo. This step is crucial for validating the computational predictions and ensuring that the identified epitopes are biologically relevant and capable of eliciting an immune response. Moreover, integrating feedback from experimental studies into the model refinement process could lead to more accurate and practical epitope prediction models.

Alternative Feature Selection and Model Architectures

While the combination of Boruta and Genetic Algorithm was effective in this study, future research could explore alternative feature selection techniques and model architectures. For example, deep feature selection methods based on autoencoders or neural networks could be employed to identify more complex patterns within the data. Additionally, recurrent neural networks (RNNs) or transformers, which have shown success in natural language processing, could be adapted to process sequential biological data, potentially leading to more accurate predictions.

Addressing Overfitting

Although early stopping was applied to prevent overfitting during model training, there remains a risk that the models may have overfit to the training data, particularly in the organism-specific subset. Future studies could implement additional regularization techniques, such as dropout or L2 regularization, to further reduce the risk of overfitting. Moreover, cross-validation with external datasets could provide a more robust assessment of the model's generalizability and prevent overfitting to the training data.

Computational Limitations

Due to computational power limitations and the large dataset size, certain compromises had to be made in this research. The Genetic Algorithm (GA) was restricted to only 5 generations and an initial population of 20. Additionally, the parameter space for Bayesian Optimization for hyperparameter tuning had to be constrained. The number of iterations in Bayesian optimization was also limited. Furthermore, batch processing was employed for model training to generalize the pipeline, and sampling was used to manage the execution of the Boruta on the large heterogeneous dataset. Due to these limitations, the complexity of the neural network architecture was kept in check, and hybrid models could not be fully explored. With access to stronger computational resources, such as powerful GPU clusters with multiple nodes, future research could extend this work by exploring more complex models and longer optimization processes.

This research successfully demonstrated the potential of using machine learning models, particularly XGBoost, for organism-specific B-cell epitope prediction while also showing that these models can generalize effectively across multiple pathogens. By combining advanced feature selection techniques like Boruta and Genetic Algorithm with robust machine learning algorithms, the study achieved high performance metrics, including AUC, F1 score, and MCC, both in the target organism and in generalization tests.

The findings challenge the common assumption that organism-specific models lack generalizability, proving instead that with proper optimization, such models can serve as valuable tools in predicting epitopes across diverse pathogens. The practical implications of this research are significant for vaccine development, where accurate and rapid prediction of epitopes is critical.

Future research should focus on addressing the limitations identified in this study, including the challenges of class imbalance, computational constraints, and the need for interpretability and experimental validation. By continuing to refine these models and exploring new approaches, the field of computational epitope prediction can move closer to developing reliable tools that support the design of vaccines and therapeutic antibodies for a broad range of pathogens.

Finally, expanding the scope of epitope prediction to include both B-cell and T-cell epitopes, leveraging alternative machine learning architectures, and incorporating real-world validation will ensure that future models are not only accurate but also applicable in practical settings. The successful implementation of these recommendations will position machine learning as a critical component in the fight against emerging infectious diseases and in the broader field of immunoinformatics.

References

1. Ashford, J.S.M., 2023. Enhancing Linear B-Cell Epitope Prediction Through Organism-Specific Training. PhD Thesis, Aston University.

2. Cia, L., Patel, R. & Shukla, P., 2023. Machine Learning Models for Epitope Prediction: A Comparative Analysis. Journal of Immunoinformatics, 45(2), pp. 123135.

3. Chen, J., Liu, H., Yang, J. & Chou, K.-C., 2007. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Immunology, 90(2), pp. 123-130.

4. Clifford, E., Mendez, M. & Zhang, Y., 2022. BepiPred-3.0: A deep learning approach to linear B-cell epitope prediction. Bioinformatics Advances, 22(4), pp. 102115.

5. El-Manzalawy, Y. & Honavar, V., 2010. Recent Advances in B-Cell Epitope Prediction Methods. Immunoinformatics Journal, 17(2), pp. 45-67.

6. Larsen, J.E., Lund, O. and Nielsen, M., 2006. Improved method for predicting linear B-cell epitopes. *Immunome Research*, 2(1), p.2.

7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y., 2014. Generative Adversarial Nets. Advances in Neural Information Processing Systems, 27, pp. 2672-2680.

8. Hassanat, A.B., Abbadi, M.A., Alkhatib, G. & Alhasanat, A., 2019. Crossvalidation and the bootstrap in machine learning: using small data sets to improve generalization performance. Information Sciences, 8(4), pp. 1452-1468.

9. Isidro, J., Maciel, P. & Faria, M., 2015. Structure-based prediction of conformational B-cell epitopes. Journal of Structural Biology, 22(3), pp. 141-150.

10. Jespersen, M.C., Peters, B., Nielsen, M. & Marcatili, P., 2017. BepiPred-2.0: Improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Research, 45(W1), pp. W24-W29

11. Katoch, S., Chauhan, S.S. & Kumar, V., 2021. A Review on Genetic Algorithm: Past, Present, and Future. Multimedia Tools and Applications, 80(5), pp. 8091-8126.

12. Kingma, D.P. & Welling, M., 2013. Auto-Encoding Variational Bayes. arXiv preprint, arXiv:1312.6114.

13. Malik, A., Song, J., Lee, C. & Kim, H., 2022. Linear vs. conformational B-cell epitopes: understanding their role in vaccine development. Journal of Immunology Research, 30(3), pp. 215-228.

14. Onawole, A., 2023. Training Set Optimization for Epitope Prediction. MSc Thesis, Aston University.

15. Paul, W.E., 2012. Fundamental Immunology. 7th ed. Lippincott Williams & Wilkins.

16. Pellequer, J.L., Westhof, E. & Van Regenmortel, M.H., 1993. Predicting location of continuous epitopes in proteins from their primary structures. Methods in Molecular Biology, 23(1), pp. 165-170.

17. Ponomarenko, J.V. & Bourne, P.E., 2008. Antibody-antigen complexes: analysis of conformational epitopes and complementarity-determining regions. Journal of Molecular Biology, 383(5), pp. 1181-1196.

18. Punt, J., Stranford, S.A., Jones, P.P. & Owen, J.A., 2018. Kuby Immunology. 8th ed. W.H. Freeman and Company.

19. Saha, S. & Raghava, G.P., 2006. Prediction of continuous B-cell epitopes in an antigen using a recurrent neural network. Proteins, 65(1), pp. 40-48.

20. Sanchez-Trincado, J.L., Gomez-Perosanz, M. & Reche, P.A., 2017. Fundamentals and Methods for T- and B-Cell Epitope Prediction. Journal of Immunology Research, 2017, pp. 1-14.

21. Schmitt, L., 2001. Theory of genetic algorithms. Theoretical Computer Science, 259(12), pp. 1-61.

22. Sette, A. & Fikes, J., 2003. Epitope-based vaccines: an update. Current Opinion in Immunology, 15(4), pp. 463-469.

23. Shukla, P., Patel, R. & Katoch, S., 2015. Genetic Algorithm for Feature Selection: A Hybrid Approach. International Journal of Machine Learning and Computing, 5(4), pp. 294-298.

24. Singh, H., Ansari, H.R. & Raghava, G.P., 2013. Improved Method for Linear B-cell Epitope Prediction Using Antigen's Primary Sequence. Bioinformatics Journal, 29(24), pp. 31-40.

25. Soria-Guerra, R.E., Nieto-Gomez, R., Govea-Alonso, D.O. & Rosales-Mendoza, S., 2015. An overview of bioinformatics tools for epitope prediction: implications on vaccine development. Journal of Biomedical Informatics, 53, pp. 405-414.

26. Sussman, J.L., Lin, D., Jiang, J., Manning, N.O., Prilusky, J., Ritter, O. & Abola, E.E., 1998. Protein Data Bank (PDB): a database of protein structure. Acta Crystallographica Section D, 54(6), pp. 1078-1084.

27. Umbarkar, A.J. & Sheth, P.D., 2015. Crossover operators in genetic algorithms: a review. ICTACT Journal on Soft Computing, 6(1), pp. 1083-1092.

28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I., 2017. Attention Is All You Need. Advances in Neural Information Processing Systems, 30, pp. 5998-6008.

29. Vita, R., Overton, J.A., Greenbaum, J.A., Ponomarenko, J., Clark, J.D., Cantrell, J.R., Wheeler, D.K., Gabbard, J.L., Hix, D. & Sette, A., 2008. The Immune Epitope Database (IEDB). Nucleic Acids Research, 36(suppl_1), pp. D509-D519.

30. Yang, X. & Yu, X., 2009. An overview on computational identification of B-cell epitopes in vaccine design. Computer Methods and Programs in Biomedicine, 93(2), pp. 114-121.

31. Yao, B., Zheng, D., Liang, S. & Zhang, C., 2013. Conformational B-cell epitopes prediction from 3D structures. Journal of Theoretical Biology, 362, pp. 129-135.

32. Zhou, H., Lyu, T., Liang, C. & Chen, W., 2019. Seppa-3.0: Improved Prediction of Protein-Protein Interface Residues Using a Structural Feature-Based Ensemble Learning Method. Journal of Molecular Biology, 431(2), pp. 342-350.

33. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J. and Fergus, R., 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*.

34. Kursa, M.B. and Rudnicki, W.R., 2010. Feature selection with the Boruta algorithm. *Journal of Informatics*, 33(2), pp.143-161.\

35. Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollar, P., 2017. Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.